Nucleic Acids Research, 2004, Vol. 32, Database issue ***D557–D559***
DOI: 10.1093/nar/gkh106

The Centre for Modeling Human Disease Gene Trap resource

Christine To[1], Trevor Epp[1,2], Tammy Reid[1,2], Qing Lan[1,2], Mei Yu[1,2], Carol Y. J. Li[1,2], Minako Ohishi[1,2], Paula Hant[1], Nora Tsao[1], Guillermo Casallo[1,3], Janet Rossant[1,4], Lucy R. Osborne[1,3,4] and William L. Stanford[1,2,5,*]

[1]Centre for Modeling Human Disease, Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada, [2]Institute of Biomaterials and Biomedical Engineering, [3]Department of Medicine, [4]Department of Molecular and Medical Genetics and [5]Institute of Medical Science, University of Toronto, Canada

Received August 15, 2003; Revised and Accepted October 13, 2003

ABSTRACT

Gene trap mutagenesis of mouse embryonic stem cells generates random loss-of-function mutations, which can be identified by a sequence tag and can often report the endogenous expression of the mutated gene. The Centre for Modeling Human Disease is performing expression- and sequence-based screens of gene trap insertions to generate new mouse mutations as a resource for the scientific community. The gene trap insertions are screened using multiplexed *in vitro* differentiation and induction assays, and sequence tags are generated to complement expression profiles. Researchers may search for insertions in genes expressed in target cell lineages, under specific *in vitro* conditions, or based upon sequence identity via an online searchable database (http://www.cmhd.ca/sub/genetrap.asp). The clones are available as a resource to researchers worldwide to help to functionally annotate the mammalian genome and will serve as a source to test candidate loci identified by phenotype-driven mutagenesis screens.

INTRODUCTION

The Centre for Modeling Human Disease (CMHD) represents a team of investigators in the greater Toronto area whose common goals are to generate mouse models of human disease and to contribute to the functional annotation of the mammalian genome. Our primary approaches are ethylnitrosourea (ENU)-induced mutagenesis of the mouse germline and gene trap vector-induced mutagenesis of embryonic stem (ES) cells (http://www.cmhd.ca/). Results from these mutagenesis approaches are organized within relational databases and are available to the research community via a user-friendly web interface.

Vectors used in the gene trapping approach contain a splice acceptor site immediately upstream of a promoterless reporter (1). Upon transcriptional activation of the endogenous *cis*-acting promoter and enhancer elements of the trapped gene, a fusion transcript is generated from the upstream coding sequence and the reporter gene, simultaneously mutating the trapped gene and reporting its expression pattern (explained in more detail, http://www.cmhd.ca/sub/genetrap/paradigm.htm). This fusion transcript also serves as a template for PCR-based gene detection strategies.

First-generation gene trap vectors primarily employed a promoterless β-*geo* reporter construct—a fusion of the β-galactosidase and neomycin resistance (*neo*) genes. Because selection requires gene expression, insertions can only be selected if the locus is active in undifferentiated ES cells, eliminating the selection of intergenic insertions but also eliminating selection of insertions in genes not expressed in undifferentiated ES cells. As with other types of stem cells, ES cells transcribe a relatively high complement of genes; however, to achieve genome-wide mutagenesis using this approach, there is a need to generate gene trap insertions independent of gene expression in undifferentiated ES cells. Furthermore, empirical evidence has shown that each gene trap vector demonstrates insertional biases. Therefore, we have developed a number of novel vectors employing different splice donors and acceptor sequences as well as different antibiotic resistance genes (http://www.cmhd.ca/sub/genetrap/vectors.htm). Recent gene trap vectors have also included recombination sites to allow recombinase-mediated removal of the antibiotic resistance cassette, as well as recombinase-mediated genetic modification or reversion of the gene trap locus. Empirical assessment of trapping efficiencies of various vector designs is an ongoing active component of the CMHD gene trap group, aiming not only to achieve better coverage of the mouse genome, but also to enable downstream characterization of the trapped locus.

*To whom correspondence should be addressed at Institute of Biomaterials and Biomedical Engineering, University of Toronto, 4 Taddle Creek Road, Room 407, Rosebrugh Building, Toronto, Ontario M5S 3G9, Canada. Tel: +1 416 946 8379; Fax: +1 416 978 4317; Email: william.stanford@utoronto.ca

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

EXPRESSION SCREENS

In addition to developing into all cell lineages and tissues *in vivo* [except trophoblast and primitive endoderm (2)], ES cells can also differentiate into a variety of cell types *in vitro* and can respond to various physiological and molecular signals (3–7). The developmental programs, gene expression profiles and cell signaling pathways of *in vitro* differentiated ES cells approximate many aspects of early embryonic development, including early organogenesis (8–11). Furthermore, ES cells differentiated *in vitro* provide a tool to observe and manipulate transient populations of cells that are intractable *in utero*. Expression- and induction-trapping exploit this use of ES cells to identify and mutate genes expressed in specific cell lineages or that respond to specific cues.

The primary goal of the CMHD gene trap project is to create a functional genomics resource in the mouse, in which genome-wide insertional mutagenesis is combined with expression screens to identify, characterize and mutate large numbers of genes involved in specific developmental and signaling pathways. Of particular interest are mesoderm-derived lineages such as hematopoietic, endothelial, vascular smooth muscle and cardiomyocyte, and genes responsive to radiation, hypoxia, retinoic acid, transforming growth factor-β (TGF-β) or D-glucosamine.

We have modified numerous ES cell *in vitro* differentiation assays into high-throughput screens. Thus, gene trap clones are grown in various *in vitro* differentiation or induction assays, then stained for *lacZ* (reporter) activity to determine an expression profile for the 'trapped' gene (http://www.cmhd.ca/sub/genetrap/expression.htm). The expression intensity, percentage of cells within a given assay expressing the reporter and images are available in the clone expression report on our website. As of August 2003, our database driven by the MySQL Relational Database Management System (http://www.msyql.com) now holds more than 7000 gene trap clones that have been screened for their *in vitro* expression. In addition, all cell-based protocols are available on our website (http://www.cmhd.ca/sub/genetrap/gtsop.htm).

GENE-DRIVEN SCREENS

Complementing expression analysis, a gene-driven screen identifies the trapped gene. We have obtained sequence information for over 1000 different gene trap clones using 5′ and/or 3′ rapid amplification of cDNA ends (RACE). For both protocols RNA is extracted from 96-well replica plates and is then subject to reverse transcription and adaptor ligation. A gene-trap-vector-specific primer together with an adaptor-specific primer are used for PCR amplification of the gene trap fusion transcript. 5′RACE is used to analyze genes trapped by all types of vectors, although it is dependent upon the activity of the endogenous promoter. As poly(A) trap vectors contain a constitutive promoter driving expression of a selectable marker that in turn captures an endogenous poly(A) signal via a splice donor signal, these vectors are also amenable to analysis by 3′RACE. 5′ and 3′RACE products are sequenced directly (Base Station sequencer, MJ Research) using vector-specific primers.

SEQUENCE PROCESSING AND IDENTIFICATION

Prudent processing of the RACE sequences is an important step towards proper analysis of the sequences. A pipeline of interactive sequence processing and analysis steps was created to facilitate trapped gene identification. First, chromatogram data from automated DNA sequencing is interpreted with the base-calling program Phred (12,13). Low-quality bases are removed using an empirically determined quality value cut-off of 11.5, which is less stringent than Phred's default value of 20.0. Then, the sequence is analyzed for the presence of the gene trap vector using direct sequence alignment (the NCBI's BlastN tool) between the vector sequence and the RACE sequence. A Perl script is then used to parse the location of the vector fragments. In the data input user interface, the identified vector regions are highlighted and sequences upstream of the vector region plus the vector region itself are removed. Sequence is also scanned for the presence of the vector splice site and RACE primer sequence by direct sequence alignment. In the event that the vector-encoding region is absent or a splicing event could not be ascertained, the sequence is disregarded. Poly(A/T) tails are removed and the resulting sequence must be at least 40 bp long for further characterization.

Prior to performing searches of the processed sequences on the public databases, any repetitive sequences are identified using RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker). Sequences are batch processed and blasted against the NCBI's non-redundant and EST databases, as well as Ensembl's cDNA, GenScan and genomic (both unmasked and masked) databases (http://www.ensembl.org/Mus_musculus/) using the NCBI's BlastN tool (14,15). Blast outputs are parsed with Perl script and imported into our database. If multiple samples corresponding to a single gene trap clone are sequenced, redundant entries are created in the database. These entries are manually curated to ensure that the results are consistent and the one with the best blast score is selected as the representative sequence tag for the clone.

If our sequence maps to an Ensembl gene, then the Ensembl database is queried to collect supplementary data such as chromosomal location, vector insertion site, specific protein domains and accession numbers (MGI, LocusLink, RefSeq and UniGene). The corresponding gene name, gene symbol, synonyms and gene ontology (GO) data are obtained from the MGI resource (http://www.informatics.jax.org/). With each new Ensembl mouse database release, we re-Blast our sequences to acquire the most up-to-date information. In addition, sequence tags are regularly being submitted to the NCBI GSS database (http://www.ncbi.nlm.nih.gov/dbGSS/) including clone data.

CUSTOM SEARCH SERVICE

The CMHD Gene Trap Database (http://www.cmhd.ca/sub/genetrap.asp) can be accessed directly or by logging in as an anonymous guest; however, a combined registration process is required to gain access to the CMHD ENU-induced mutant mouse database and/or to utilize the CMHD Gene Trap Database custom search service. Registration is freely available to all accredited scientific researchers. Researchers using the free custom search service can submit their choice of

Unigene and/or MGI accession numbers, DNA sequences (FASTA formatted), protein domain accession numbers (Interpro ID), GO accession numbers and/or chromosomal locations as search terms. These search terms form a user database, which will be used to query the Gene Trap Database as new clones are added. Users will receive email notification of any new clones matching their search criteria.

WEB INTERFACE

A user-friendly web-based interface allows access to the CMHD Gene Trap Database (http://www.cmhd.ca/sub/genetrap.asp) and facilitates both sequence and expression-based queries. Expression-based queries are made using a series of drop-down lists connected by Boolean operators. Thus, users can search for clones exhibiting any combination of reporter gene expression in undifferentiated ES cells and a variety of differentiated cell types. A separate field retrieves clones up- or down-regulated in a number of different induction screens. Clones are screened for response to radiation, hypoxia, TGF-β and more recently, for D-glucosamine.

Search results are displayed as a list of gene trap clone IDs, which are hyperlinked to individual clone expression reports. These expression reports list the gene trap vector used, results from *in vitro* differentiation and induction screens, often including, when positive *lacZ* expression is observed, an image file. The clone expression report page also contains a hyperlink to the corresponding sequence report, containing the processed RACE tag sequence and associated annotation.

A separate page is available for sequence-based queries. Users can paste their DNA sequences of interest in FASTA format in the designated query box to perform a BlastN search of our sequence database. The Blast output page has links to the clone sequence and expression reports. Users can also query our database by gene name, gene symbol, chromosomal location, protein domains, gene ontology annotation and accession number via our web interface.

ES cell clones are available to the scientific community unencumbered by intellectual property claims. Requests are assessed a nominal charge, set by the International Gene Trap Consortium (IGTC) (16) (http://www.igtc.ca/), to cover the expenses of cryo-storage, thawing, restocking and shipping. Links to the gene trap labs that are a part of the IGTC, which use complementary vectors to our own, can be found on our links page (http://www.cmhd.ca/sub/CMHDlinks.htm#GT).

ACKNOWLEDGEMENTS

This work has been funded by grants from the Canadian Institutes of Health Research Special Collaborative Genomics Program, Genome Canada through a grant from the Ontario Genomics Institute, and the National Institutes of Health, USA.

REFERENCES

1. Cordes,S.P., Cohn,J.B. and Stanford,W.L. (2001) Gene Trap mutagenesis: past, present and beyond. *Nature Rev. Genet.*, **2**, 756–768.
2. Beddington,R.S. and Robertson,E.J. (1989) An assessment of the developmental potential of embryonic stem cells in the midgestion mouse embryo. *Development*, **105**, 733–737.
3. Tropepe,V., Hitoshi,S., Sirard,C., Mak,T.W., Rossant,J. and van der Kooy,D. (2001) Direct neural fate specification from embryonic stem cells: a primitive mammalian neural stem cell stage acquired through a default mechanism. *Neuron*, **30**, 65–78.
4. Doetschman,T.C., Eistetter,H., Katz M., Schmidt,W. and Kemler,R. (1985) The *in vitro* development of blastocyst-derived embryonic stem cell lines: formation of visceral yolk sac, blood islands and myocardium. *J. Embryol. Exp. Morphol.*, **87**, 27–45.
5. Wang,R., Clark,R. and Bautch,V.L. (1992) Embryonic stem cell-derived cystic embryoid bodies form vascular channels: an *in vitro* model of blood vessel development. *Development*, **114**, 303–316.
6. Nakano,T., Kodama,H. and Honjo,T. (1994) Generation of lymphohematopoietic cells from embryonic stem cells in culture. *Science*, **265**, 1098–1101.
7. Choi,K., Kennedy,M., Kazarov,A., Papadimitriou,J.C. and Keller,G. (1998) A common precursor for hematopoietic and endothelial cells. *Development*, **125**, 725–732.
8. Schmitt,R.M., Bruyns,E. and Snodgrass,H.R. (1991) Hematopoietic development of embryonic stem cells *in vitro*: cytokine and receptor gene expression. *Genes Dev.*, **5**, 728–740.
9. Bautch,V.L., Stanford,W.L., Rapoport,R., Russell,S., Byrum,R.S. and Futch,T.A. (1996) Blood island formation in attached cultures of murine embryonic stem cells. *Dev. Dyn.*, **205**, 1–12.
10. Abe,K., Niwa,H., Iwase,K., Takiguchi,M., Mori,M., Abe,S.I., Abe,K. and Yamamura,K.I. (1996) Endoderm-specific gene expression in embryonic stem cells differentiated to embryoid bodies. *Exp. Cell Res.*, **229**, 27–34.
11. Schuldiner,M., Yanuka,O., Itskovitz-Eldor,J., Melton,D.A. and Benvenisty,N. (2000) Effects of eight growth factors on the differentiation of cells derived from human embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **97**, 11307–11312.
12. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated squencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.
13. Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
14. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Nadeau,J.H., Balling,R., Barsh,G., Beier,D., Brown,D.M., Bucan,M., Camper,S., Carlson,G., Copeland,N. *et al.* (2001) Annotating genome sequences with biological functions in mice. *Science*, **291**, 1251–1255.